# Speech/Music Discrimination via Energy Density Analysis

Stanisław Kacprzak and Mariusz Ziółko

Department of Electronics, AGH University of Science and Technology
al. Mickiewicza 30, Kraków, Poland
{skacprza, ziolko}@agh.edu.pl

.

**Abstract.** In this paper we suggest to apply a new feature, called Minimum Energy Density (MED), in discrimination of audio signals between speech and music. Our method is based on the analysis of local energy for 1 or 2.5 seconds audio signals. An elementary analysis of the power distribution is an effective tool supporting the decision making system. We compare our feature with Percentage of Low Energy Frames (LEF), Modified Low Energy Ratio (MLER) and examine their efficiency for two separate speech/music corpora.

**Keywords:** speech/music discrimination, sound classification, audio content analysis

## 1 Introduction

Discrimination between speech and music has applications in different areas of speech processing, such as voice activity detection (VAD), automatic corpus creation [11] and as part of modern hearing aids [1]. For the purpose of this discrimination many features, in time as well as in frequency domain, have been proposed [2], [9]. The most common are: 4 Hz modulation energy, entropy modulation, spectral centroid, spectral flux, zero-crossing rate and cepstral coefficients, but more complex parameters like wavelet-based parameters [3] are also explored. Recognition rate over 98% [8], [9], has been reported for subsets of these features and their variations. Current research is focused on achieving high recognition rate with aspect of minimizing required computations. In this paper we focus on speech/music discrimination based on energy features. We analyse energy distribution in speech and music signals and upon this analysis we introduce a new feature Minimum Energy Density (MED). We compare this feature with Percentage of Low Energy Frames (LEF), Modified Low Energy Ratio (MLER) and examine their efficiency for corpus collected by Scheirer and Slaney [9] and a second one, created by us.

Kacprzak S., Ziółko M.: Speech/Music Discrimination via Energy Density Analysis, Proceedings of the 1st International Conference on Statistical Language and Speech Processing, Tarragona 2013, Springer pp. 135-142.
The final publication is available at link.springer.com

## 2 Energy Features

It is very intuitive to try to discriminate speech and music based on shape of signal's energy envelope. As Fig. 1 shows, speech signal has characteristic high and low amplitude parts, which represent voiced and unvoiced speech, respectively. On the other hand, the envelope of music signal is more steady. Moreover, we know that speech has a characteristic 4 Hz energy modulation, which matches the syllabic rate [9].

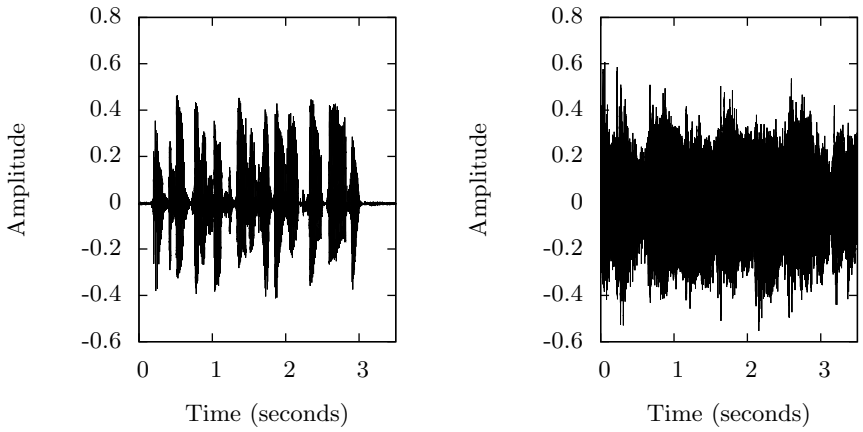


**Fig. 1.** Speech (left) and music (right) samples

Saunders [8] stated: "The energy contour is well known to be capable of separating speech from music." His algorithm however, was based on zero-crossings rate features and 90% accuracy was reported. It is interesting that after adding a new feature, which was a measure of energy minima below some threshold relative to peak energy, accuracy rose to 98%. Results based only on this energy feature were not presented. Measure of rapid changes in speech signal was the base of speech/music discrimination in hardware device described in patent [4].

In [9] authors define Percentage of Low Energy Frames (LEF) feature as percentage of frames within 1 s window with root mean square (RMS) power below 50% of window mean RMS power. This feature alone provides 14% error rate and was the fastest one in the sense of computational time. Similar feature was proposed in [5], but authors used short term energy instead of RMS. Wang, Gao and Ying in [10] explore this idea by introducing Modified Low Energy Ratio (MLER) which is different from LEF in the fact that percentage of the window mean short term energy is not fixed to 50%, but its value is subject to change.

The formal definition of MLER [10] is

$$MLER = \frac{1}{2N} \sum_{n=1}^{N} \left[ \operatorname{sgn} \left( lowthres - E(n) \right) + 1 \right], \tag{1}$$

where

$$lowthres = \delta \cdot \sum_{n=1}^{N} E(n) \tag{2}$$

and $N$ is the total number of frames in a window, $E(n)$ is frame short time energy and $\delta$ is control coefficient.

These features take under consideration skewness of energy distribution in speech (Fig. 2), caused by the fact that there are many low energy or quiet frames in speech, also more than in music. However, these features ignore energy distribution within a window. Thus, they will fail in the presence of speech window with low mean energy, that can appear for example when a fricative is followed by a pause, or in case of whole silent window, which may occur if the person is speaking slowly. Moreover, because of relative character of this feature, MLER can fail in the presence of additive noise, since it would be necessary to increase $\delta$ with the decrease of SNR.

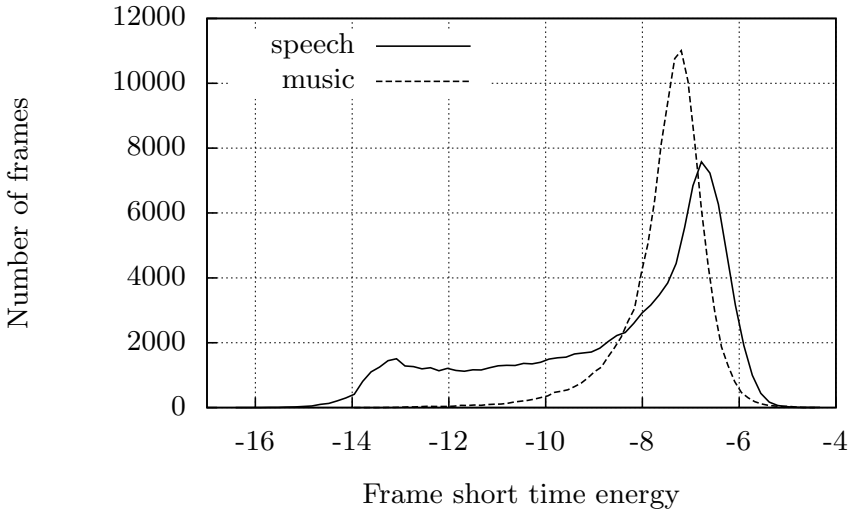


**Fig. 2.** Histogram outlines of normalized short time energy calculated for audio samples used in [9]. Values of energy have been log-transformed

The number of energy dips below the value of threshold, which is little above noise level, was used as a feature in [7], where 86% accuracy was reported for 5 s

windows, but tests where performed on very rigorous music data which contained single instrument music. Our feature explores idea of classification based on energy dips.

## 3 Minimum Energy Density Feature

We know from energy distribution (Fig. 2) that speech has more low energy frames than music. We also know that speech has 4 Hz energy modulation, which implies four energy minima in 1 s window. These facts allow us to suspect that the presence of the frame with energy below some calculated threshold is sufficient to distinguish between speech and music. The disadvantage of this approach is inability to rely on some fixed threshold value, due to differences in signals power. To overcome that, we calculate distribution of short time frame energy inside some time window, which we refer to as *normalization window*. Normalization window has to be long enough to capture the nature of the signal. For example 1 s window seems a bad idea, since in case of window containing breathe pause we would get distribution close to uniform and information about low energy of that window would be lost. We define normalized short time frame energy as

$$\bar{E}(n) = \frac{E(n)}{\sum_{k=1}^{N} E(k)} \ . \tag{3}$$

Next step is to find minimum $\bar{E}(n)$ in the *classification window*. Length of the classification window can be shorter than the length of normalization window and it defines classification resolution. Taking into account the 4 Hz energy modulation characteristic for speech, the length of the classification window should be at least 250 ms. We define Minimum Energy Density (MED) for $k$-th classification window as

$$MED(k) = \min\{\bar{E}(n) : (k-1) \cdot M + 1 \leq n \leq k \cdot M\}, \tag{4}$$

where $M$ is the number of frames in the classification window.

During training phase a threshold value for MED is found so that the windows with MED below that threshold are classified as speech and the rest as music. In fact, for classifying unseen data, there is no need to find minimum value of the classification window as in (4), because finding any frame with energy below the threshold is sufficient to classify the window as speech. Additionally we can reduce needed computations by, instead of normalizing each frame in normalization window, scaling the threshold. Final decision about class for a classification window is given by

$$class(k) = \begin{cases} speech & \text{if} \quad \exists n : E(n) < \lambda, \ \text{ where } \ (k-1) \cdot M + 1 \leq n \leq k \cdot M \\ music & \text{otherwise} \end{cases}, \tag{5}$$

where

$$\lambda = threshold \cdot \sum_{n=1}^{N} E(n) \ . \tag{6}$$

Figure 3 shows histogram outlines of MED feature for speech and music signals.

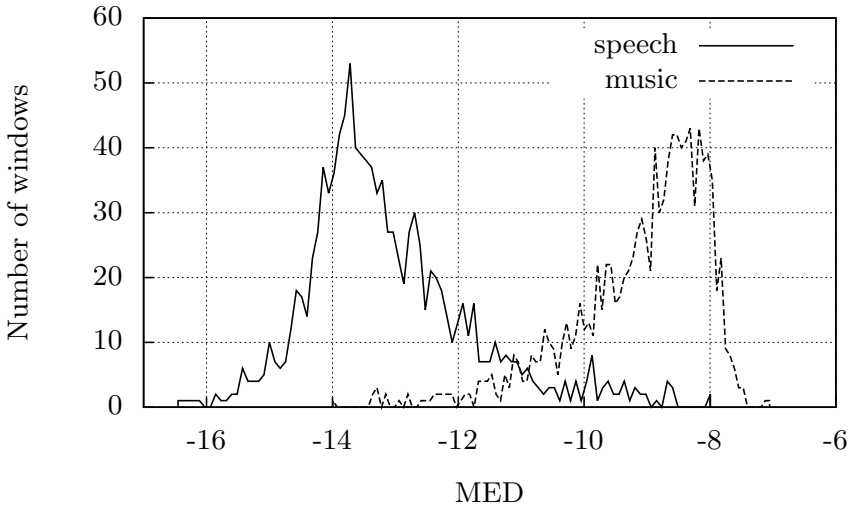


**Fig. 3.** Histogram outlines of MED calculated on audio samples used in [9]. Values of MED have been log-transformed

## 4 Test Corpora

To evaluate our algorithm we use two separate audio data sets. First set, which will be referred to as A, is the same that was used in [9] and consists of eighty 15-second long audio samples of speech and the same amount of music samples. As authors stated, the data was collected by digitally sampling FM tuner (16-bit at a 22.05 kHz sampling rate). Speech data contains male and female speakers, in quiet and in noisy conditions. Music data set contains variety of music styles, with and without vocals. The second data set, which will be referred to as B, was collected by us. We also prepared eighty 15-second long audio samples of speech from mp3 of Polish audio-books and same amount of music derived from private mp3 library (16-bit at a 44.1 kHz, stereo files were transformed to mono). The speech samples feature both male and female, mostly professional speakers and actors while in music data set we try to capture variety of music genres like rock, pop, jazz, dance and reggae.

## 5 Experiment Evaluation

We examine our algorithm using 10 ms frames, 15 s (whole audio sample) normalization window and 1 s and 2.5 s classification windows. We compare results of our new feature with LEF and MLER. For MLER we analyse the effect of $\delta$ value first. The results, which are shown in Fig. 4, imply that in our case $\delta = 0.1$, as suggested in [10], is not the best possible option. Instead we choose

$\delta = 0.02$, which is the cross point of lines representing average accuracy. To evaluate our algorithms for every experiment we use over 10 cross-validation runs. In each run we calculate MED for all samples. 70% of calculated parameters selected at random were used as training set and the remaining 30% were used for testing. During the training session the best threshold value that maximizes overall classification accuracy over the training set was found and that threshold was used to classify data under test set. The mean results of cross-validation runs of speech/music discrimination for 1 s and 2.5 s classification windows are shown in Tab. 1 and Tab. 2, respectively.

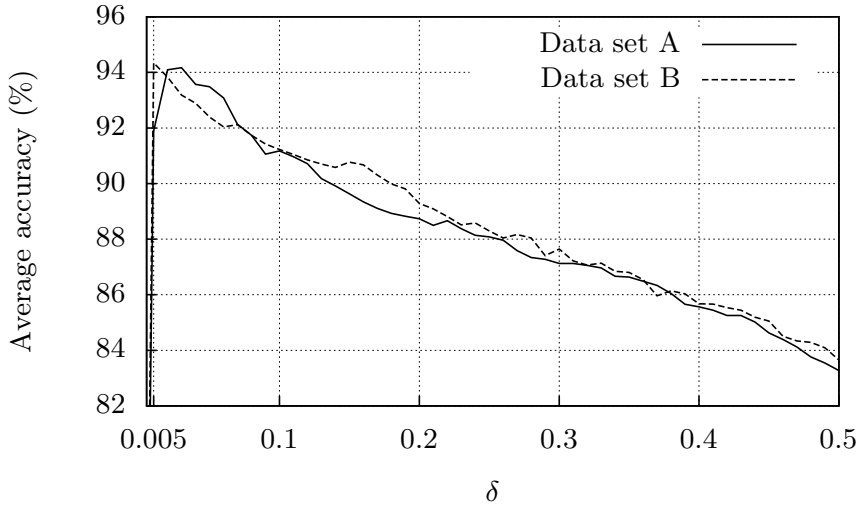


**Fig. 4.** Average accuracy of correct recognition based on MLER in function of parameter $\delta$.

**Table 1.** Correct classification results (mean and standard deviation) for the 1 s classification window

| | Data set A | | | Data set B | | |
|---|---|---|---|---|---|---|
| | LEF | MLER | MED | LEF | MLER | MED |
| speech | 87.5 ± 3.9% | 91.3 ± 1.0% | 91.6 ± 1.5% | 88.1 ± 2.3% | 95.1 ± 1.2% | 94.9 ± 1.3% |
| music | 90.1 ± 3.2% | 96.7 ± 0.6% | 95.3 ± 1.4% | 90.4 ± 1.3% | 92.6 ± 1.4% | 95.3 ± 1.0% |
| **total** | **88.8 ± 0.9%** | **94.0 ± 0.3%** | **93.5 ± 0.4%** | **89.3 ± 1.3%** | **93.8 ± 0.6%** | **95.1 ± 0.6%** |

**Table 2.** Correct classification results (mean and standard deviation) for the 2.5 s classification window

| | Data set A | | | Data set B | | |
|---|---|---|---|---|---|---|
| | LEF | MLER | MED | LEF | MLER | MED |
| speech | 92.4 ± 2.5% | 95.4 ± 2.1% | 94.5 ± 2.2% | 96.3 ± 1.7% | 96.8 ± 2.3% | 98.0 ± 1.1% |
| music | 91.0 ± 3.5% | 95.7 ± 1.6% | 97.0 ± 1.4% | 94.3 ± 1.7% | 95.9 ± 2.0% | 96.0 ± 1.5% |
| **total** | **91.7 ± 1.6%** | **95.5 ± 1.2%** | **95.8 ± 1.5%** | **95.3 ± 1.1%** | **96.3 ± 1.2%** | **97.0 ± 0.7%** |

## 6 Conclusions

The results in Tab. 1 and Tab. 2 demonstrate that MED method performs better than LEF and slightly better or similarly as MLER. However, our method is not dependent on any parameter, like $\delta$ in case of MLER, that has a strong effect on accuracy and its optimal value depends on tested data. In case of the 2.5 s classification window our method achieves 95.8% accuracy for data set A and 97% accuracy for data set B, what are very high results for single feature. In contrast, in [9] authors reported 98.6% accuracy on the 2.4 s window using GMM classifier based on 3 features.

Furthermore, in case of our algorithm, after finding the frame with energy below the threshold, the calculation stops for a given window, resulting in the reduction of the expected number of calculations. This fact and the manner in which the threshold energy value based on MED is found, distinguishes our algorithm from one presented in [7] and shows that MED is sufficient for good discrimination in case of speech and typical modern music. Considering its good performance and low computation load, the algorithm which is based on MED feature allows more effective speech/music discrimination.

It needs to be pointed out that our tests include only recordings of speech or music. There were no examples of speech over music, which would imply three class discrimination, because classifying such signal as speech or music is subjective. Nevertheless, our method alone has potential to be used for tasks like automatic corpus [6] creation from sources for which we have prior knowledge that are compound of alternating speech and music like audio-books, language courses or radio drama.

## Acknowledgements

The project was funded by the National Science Centre allocated on the basis of a decision DEC-2011/03/B/ST7/00442.

## References

1. Cabañas Molero, P., Ruiz Reyes, N., Vera Candeas, P., Maldonado Bascon, S.: Low-complexity f0-based speech/nonspeech discrimination approach for digital hearing aids. Multimedia Tools and Applications 54, 291–319 (2011)

2. Carey, M., Parris, E., Lloyd-Thomas, H.: A comparison of features for speech, music discrimination. In: Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on. vol. 1, pp. 149 –152 vol.1 (mar 1999)
3. Didiot, E., Illina, I., Fohr, D., Mella, O.: A wavelet-based parameterization for speech/music discrimination. Comput. Speech Lang. 24(2), 341–357 (Apr 2010)
4. Jones, R.C.: Electronic device for automatically discriminating between speech and music forms (1956), US Patent 2761897
5. Lu, L., Jiang, H., Zhang, H.: A robust audio classification and segmentation method. In: Proceedings of the ninth ACM international conference on Multimedia. pp. 203–211. MULTIMEDIA '01, ACM, New York, NY, USA (2001)
6. Masior, M., Ziółko, M., Kacprzak, S.: Multi-lingual speech samples base. URL: `http://speechsamples.agh.edu.pl/`
7. Okamura, S., Aoyama, K.: An experimental study of energy dips for speech and music. Pattern Recognition 16(2), 163–166 (1983)
8. Saunders, J.: Real-time discrimination of broadcast speech/music. In: Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on. vol. 2, pp. 993 –996 vol. 2 (may 1996)
9. Scheirer, E., Slaney, M.: Construction and evaluation of a robust multifeature speech/music discriminator. In: Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on. vol. 2, pp. 1331 –1334 vol.2 (apr 1997)
10. Wang, W., Gao, W., Ying, D.: A fast and robust speech/music discrimination approach. In: Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on. vol. 3, pp. 1325 – 1329 vol.3 (dec 2003)
11. Wei, Z., Ranran, D., Minhui, P., Qiuhong, W.: Automatic speech corpus construction from broadcasting speech databases. In: Computational Intelligence and Security (CIS), 2010 International Conference on. pp. 639 –643 (dec 2010)